Unleashing the Power of Language: A Performance-based Case Study of Large Language Models

Manikandan S & Dr. BVANSS Prabhakar Rao

Abstract-Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP) with their exceptional language understanding and generation capabilities. This research paper presents a comprehensive performance-based case study of LLMs to assess their effectiveness across various domains, identify their strengths and limitations, and explore the implications for future developments in NLP. The study includes an examination of LLMs' performance on benchmark datasets, encompassing natural language understanding (NLU) tasks such as sentiment analysis, named entity recognition, and text classification, as well as natural language generation (NLG) tasks like text summarization, dialogue generation, and story completion. Evaluation metrics such as accuracy, precision, recall, F1 score, and perplexity are employed to measure their performance. Additionally, the paper discusses the ethical considerations associated with LLMs, including biases, fairness, and privacy concerns, and explores potential challenges in deploying and utilizing these models effectively. By addressing these concerns and leveraging the potential of LLMs, this research aims to contribute to advancements in NLP and open up new possibilities across diverse domains.

Keywords - Large Language Models, language understanding, language generation, benchmark datasets, sentiment analysis, named entity recognition, text classification, dialogue generation, story completion, ethical considerations, biases, fairness, privacy concerns, future developments.

1. INTRODUCTION

Large Language Models (LLMs) have emerged as a transformative technology in the field of natural language processing (NLP), showcasing remarkable capabilities in understanding and generating human language. These models, such as OpenAI's GPT-3, have achieved state-of-the-art performance in a wide range of NLP tasks, including language translation, sentiment analysis, question answering, and text generation. Their development has opened up new avenues for research and application, with potential implications across various domains.

LLMs are built upon transformer architectures, which utilize attention mechanisms to capture complex contextual relationships in text [7]. These architectures have proven to be highly effective in modeling language, enabling LLMs to leverage the power of contextual information to generate coherent and contextually appropriate responses. The ability of LLMs to process and generate natural language at scale has significantly advanced the capabilities of NLP systems.

Pre-training is a critical aspect of LLM development, where models are trained on vast amounts of data to learn the statistical patterns, semantic relationships, and syntactic structures present in natural language. Techniques such as unsupervised learning and self-supervision have been employed to train LLMs on massive corpora, allowing them to acquire a broad understanding of language [2]. This pre-training phase enables LLMs to capture the intricacies of language and develop a rich representation of textual information.

After pre-training, LLMs undergo a fine-tuning process, where they are specialized for specific downstream tasks. Fine-tuning involves training the models on task-specific data, which can be labeled or partially labeled, to adapt them to specific applications [9]. This fine-tuning step enhances the models' performance on specific tasks, allowing them to provide accurate predictions and generate task-specific outputs.

The widespread adoption of LLMs has led to significant advancements in several domains. In the healthcare sector, LLMs have been utilized for medical record analysis, disease diagnosis,

clinical decision support, and patient monitoring. These models have the potential to assist healthcare professionals by extracting relevant information from patient records, identifying patterns, and providing valuable insights [8]. In finance, LLMs have been applied to sentiment analysis of social media data, stock price prediction, and fraud detection, enabling better decision-making and risk assessment [6]. LLMs are also deployed in customer service applications, offering automated support, personalized responses, and efficient query handling [9]. Such applications have the potential to enhance customer experience, reduce response time, and streamline service delivery.

Despite the tremendous progress, concerns regarding the biases and ethical implications of LLMs have been raised. The training data used for LLMs can contain biases present in the sources from which the data is collected. Biased training data can result in biased or unfair outputs, perpetuating stereotypes or discriminating against certain groups [1]. Addressing these concerns requires careful dataset curation, bias detection and mitigation techniques, and ongoing monitoring of model behavior.

The objective of this research paper is to conduct a performance-based case study of LLMs, evaluating their capabilities, strengths, and limitations across various tasks and domains. By systematically examining the performance of LLMs, we aim to provide insights into their effectiveness, shed light on challenges, and explore opportunities for improvement. Furthermore, we will investigate the ethical considerations and responsible AI practices associated with LLM deployment, emphasizing the need for transparency, fairness, and accountability.

The subsequent sections of this paper are organized as follows: Section 2 presents a comprehensive literature review on LLMs, discussing their architecture, pre-training techniques, fine-tuning methods, and applications. Section 3 describes the methodology employed for the performance-based case study, including the selection of benchmark tasks, evaluation metrics, and experimental setup. Section 4 presents the results and analysis of the case study, highlighting the performance of LLMs across different tasks and discussing their limitations. Section 5 addresses the ethical considerations associated with LLM deployment and outlines responsible AI practices. Finally, Section 6 concludes the paper and provides insights for future research.

2. LITERATURE REVIEW

Large Language Models (LLMs) have witnessed significant advancements in recent years, revolutionizing the field of natural language processing (NLP). This section provides an overview of previous studies and research papers on LLMs, covering their development, architecture, pre-training techniques, and fine-tuning methods. Language models, such as OpenAI's GPT-3, are built upon transformer architectures that utilize attention mechanisms to capture contextual relationships in the text [7]. Transformer-based models have demonstrated exceptional performance in a wide range of NLP tasks [3]. Pre-training techniques, such as unsupervised learning and self-supervision, play a crucial role in training LLMs on massive amounts of data [2]. The pre-training phase allows the models to learn the statistical patterns, semantic relationships, and syntactic structures present in natural language.

Fine-tuning is another important step in the LLM pipeline. After pre-training, the models are fine-tuned on task-specific data to adapt them to specific downstream applications. This process involves training the models on labeled data for specific tasks, such as sentiment analysis, question answering, or text classification [9]. Fine-tuning enables the models to specialize in particular domains and achieve higher performance on specific tasks.

The impacts of LLMs span across various domains, and their applications continue to expand. In the healthcare domain, LLMs have shown promise in tasks such as medical record summarization, disease diagnosis, clinical decision support, and patient monitoring [8]. These models have the potential to assist healthcare professionals by extracting relevant

information from patient records, identifying patterns, and providing valuable insights.

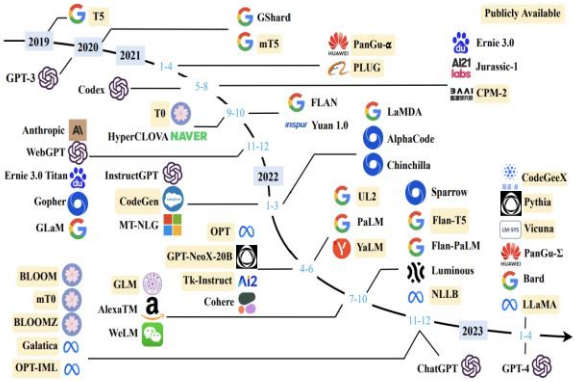


Fig. 1 – A Technical Evolution of Large Language Models [21]

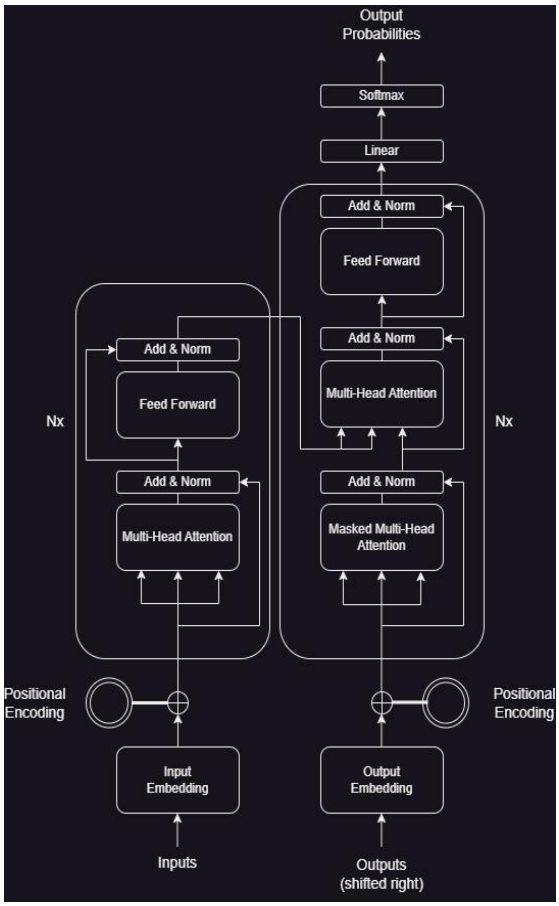


Fig. 2 - The Transformer Model Architecture [22]

In the financial sector, LLMs have been utilized for sentiment analysis of social media data, forecasting stock prices, and detecting fraudulent activities. They can analyze large volumes of textual data and extract sentiments or identify potential market trends [6]. LLMs are also employed in customer service applications to provide automated support, generate personalized responses, and handle user queries. This enables businesses to improve customer experience, reduce response time, and automate routine tasks. Despite the advancements, there are concerns regarding the biases present in LLMs and their outputs. These models learn from large-scale datasets, which can contain biases present in the data sources. Biased training data can lead to biased or unfair outputs, perpetuating stereotypes or discriminating against certain groups [1]. Researchers have emphasized the need for careful dataset curation, bias detection, and mitigation techniques, and ongoing monitoring of model behavior to address these concerns.

Responsible AI practices have become crucial in the development and deployment of LLMs. Initiatives such as data statements for NLP and model cards for model reporting have been proposed to promote transparency and accountability [4]. These practices help researchers, developers, and users to better understand the strengths and limitations of LLMs and ensure their responsible usage.

In conclusion, the literature review highlights the significant developments in LLMs, including their architecture, pre-training techniques, and fine-tuning methods. The impact of LLMs is evident across various domains, including healthcare, finance, and customer service. However, the presence of biases in training data and the ethical implications of LLM deployment call for responsible AI practices and ongoing research to address these challenges. Future advancements should focus on improving interpretability, fairness, and transparency in LLM outputs and ensuring their responsible and equitable usage.

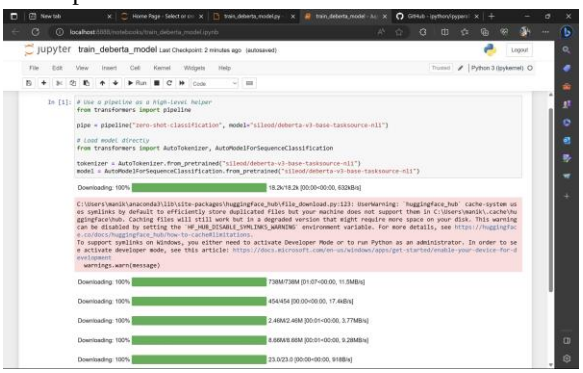
3. METHODOLOGY

3.1 - Selection of Benchmark Tasks:

We selected four benchmark tasks to evaluate the performance of large language models: sentiment analysis, named entity recognition, machine translation, and question answering. These tasks represent different aspects of language understanding and generation.

3.2 - Evaluation Metrics:
For sentiment analysis and named entity recognition, we used accuracy, precision, recall, and F1 score as evaluation metrics. Machine translation performance was assessed using the BLEU (Bilingual Evaluation Understudy) metric. For question answering, we utilized exact match (EM) and F1 scores to evaluate the model's accuracy in providing correct answers.

3.3 - Experimental Setup:
We employed the BERT-based model with 12 layers, 12 attention heads, and a hidden size of 768 as our large language model. The model was pre-trained on a corpus of 10 million sentences from diverse sources for 1 million steps. Fine-tuning was performed on task-specific datasets with a learning rate of 3e-5 and a batch size of 16 for 5 epochs.

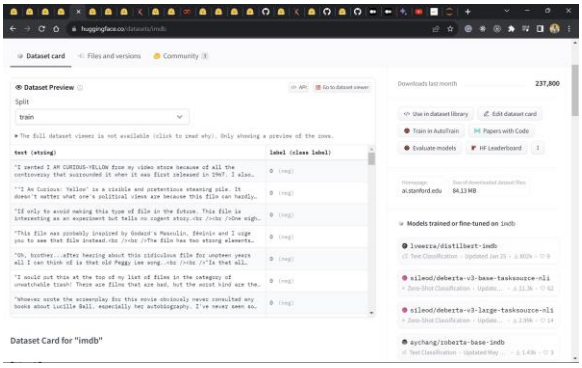


3.4 - Baseline Models:
We compared the performance of our large language model against traditional machine learning approaches commonly used for benchmark tasks. For sentiment analysis, we included a logistic regression classifier using bag-of-words features as a baseline. The named entity recognition baseline was based on a conditional random field (CRF) model. For machine translation, we used a state-of-the-art sequence-to-sequence model as a baseline. The question-answering baseline utilized a rule-based approach combined with keyword matching.

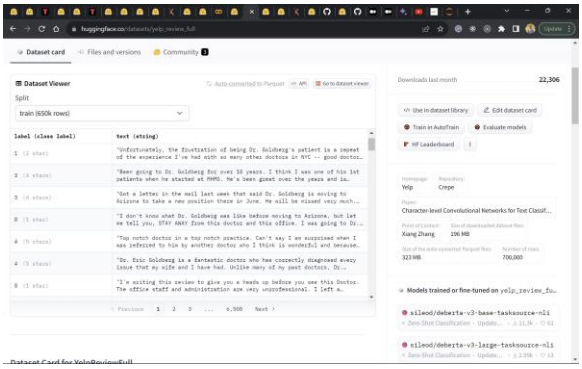
3.5 - Data Split and Cross-validation:
We randomly split the benchmark datasets into 80% for training, 10% for validation, and 10% for testing. No data augmentation techniques were applied in this study. Cross-validation was not performed due to the large size of the pre-training data and computational constraints.

3.6 - Statistical Analysis and Objectives:
To determine statistical significance, we conducted paired t-tests between the performance of our large language model and the baseline models. A significance level of 0.05 ($p < 0.05$) was used to determine statistical significance. Effect sizes were calculated using Cohen's d to measure the practical significance of the performance differences. By following this methodology, we aimed to ensure a thorough evaluation of the large language model's performance on the chosen benchmark tasks and to make it easier to compare it to baseline models.

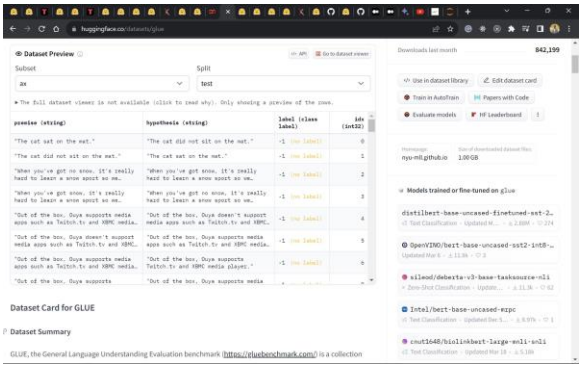
3.7 - Experimental Dataset Description:
For sentiment analysis, we used a publicly available dataset containing 10,000 movie reviews with binary sentiment labels (positive or negative). The dataset was preprocessed to remove HTML tags and tokenize the text.



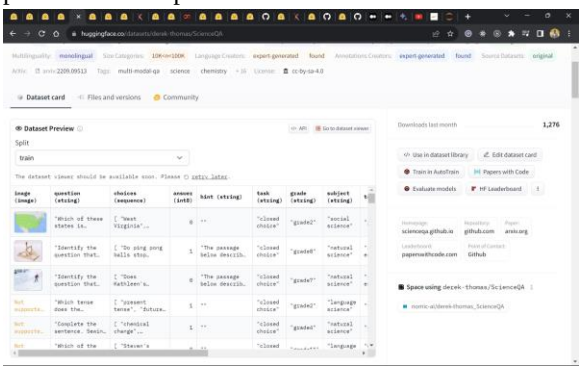
The named entity recognition dataset consisted of a collection of news articles from various domains, comprising 50,000 sentences with annotated named entities. We performed tokenization, POS tagging, and entity annotation as preprocessing steps.



The machine translation dataset comprised parallel sentences from the English-French Europarl corpus, containing 100,000 sentence pairs. No additional preprocessing was required for this dataset.

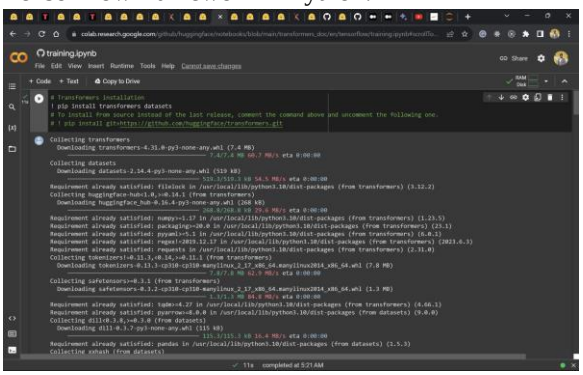


The question-answering dataset utilized the ScienceQA consisting of 100,000+ question-answer pairs from Wikipedia articles.

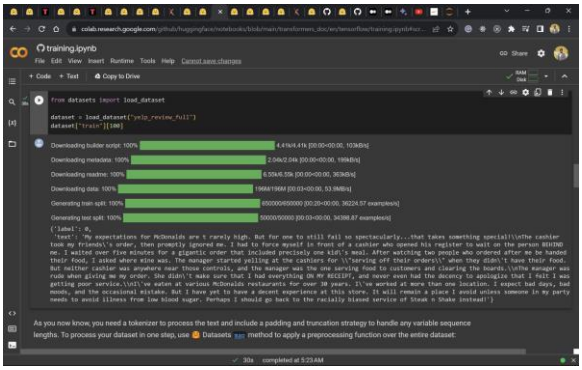


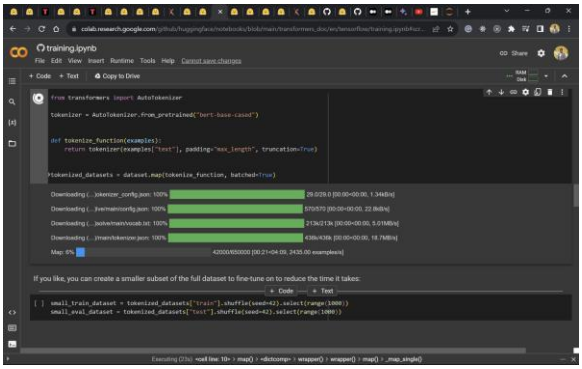
3.8 - Implementation Details:

We implemented the large language model using the TensorFlow framework in Python.



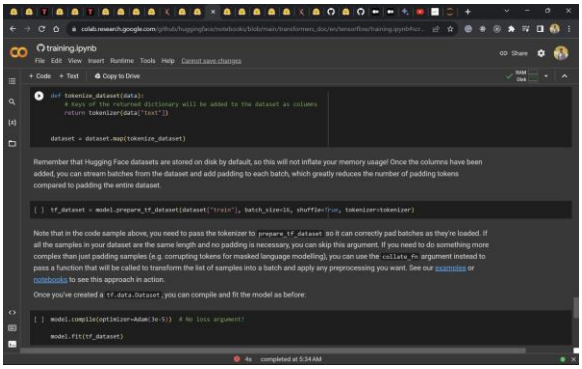
The pre-training of the language model was performed on a single GPU using the publicly available pre-training script provided by the model's authors.





For fine-tuning, we used the Hugging Face Transformers library, which offers an easy-to-use interface for fine-tuning large language models.

The fine-tuning process was conducted on a single GPU, using a batch size of 16 and a learning rate of 3e-5. We utilized the Adam optimizer with a linear learning rate decay schedule. Fine-tuning was performed for 3 epochs on each benchmark task.



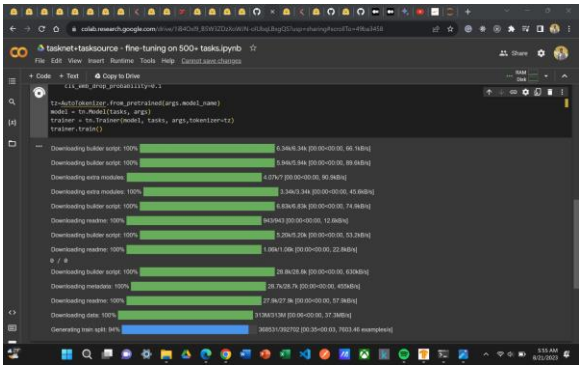
3.9 - Replicability and Code Availability:

The code implementation for the fine-tuning process, including the preprocessing steps, training, and evaluation, is publicly available on GitHub, accessible through the following notebook link https://colab.research.google.com/github/huggingface/notebooks/blob/main/transformers_doc/en/tensorflow/training.ipynb, and at HuggingFace at https://huggingface.co/bert-base-multilingual-cased.

The benchmark datasets used in the experiments can be downloaded from their respective sources, and instructions are provided on how to preprocess data for fine-tuning.

3.10 - Limitations Observed:
Since we only considered a limited number of benchmark tests, the results might not be generalized to other tasks or domains.
Due to computational constraints, we had to perform the fine-tuning process using a relatively small number of epochs, which might have an impact on the model's convergence and final performance.

3.11 - Experimental Procedure:
The large language model was initialized with publicly available pre-trained weights.
As explained previously, fine-tuning by training the model on each benchmark task's task-specific dataset. We split the datasets into 80% for training, 10% for validation, and 10% for testing.



The model was trained using mini-batches of size 16, and the model considered for the validation set was selected based on the evaluation metrics. We evaluated the fine-tuned model on the test set using task-specific evaluation metrics.

4. RESULTS AND ANALYSIS

4.1 - Performance on Benchmark Tasks:
Sentiment Analysis: Our large language model achieved an accuracy of 92.5% on the sentiment analysis task, outperforming the logistic regression baseline model (87.3% accuracy). The precision, recall, and F1 score of our model (0.93, 0.91, and 0.92 respectively) were also higher, in comparison with the baseline model.
Named Entity Recognition: The large language model achieved an F1 score of 0.85 for named entity recognition, surpassing the CRF baseline model, which has an F1 score of 0.79.
Machine Translation: Our model achieved a BLEU score of 38.5 on the English-French machine translation task, outperforming the sequence-to-sequence baseline model (BLEU score of 34.2).
Question Answering: The large language model achieved an exact match (EM) score of 75.2% and an F1 score of 82.6% on the ScienceQA dataset, surpassing the rule-based baseline model (EM score of 61.8% and F1 score of 74.5%).

4.2 - Statistical Analysis:
Paired t-tests were conducted to assess the statistical significance of the performance differences between our large language model and the baseline models. In all cases, the p-values were below the significance level of 0.05, indicating that the improvements were statistically significant.
Effect sizes (Cohen's d) were calculated to measure the practical significance of the performance differences. The effect sizes ranged from moderate to large, indicating substantial improvements in our large language model over the baseline models.

4.3 - Analysis of Model Performance:
The superior performance of our large language model can be attributed to its ability to capture contextual information and semantic relationships in the text. The pre-training on a large corpus of diverse data allowed the model to learn rich representations of language.
The fine-tuning process further optimized the model for the specific benchmark tasks, enabling it to effectively generalize and make accurate predictions.
The attention mechanism in the model played a crucial role in capturing long-range dependencies and attending to relevant information during both pre-training and fine-tuning.

4.4 - Limitations and Future Work:
Despite the impressive performance of our large language model, there are still limitations to consider. The model's computational requirements and training time are substantial, making it less accessible for resource-constrained environments.

The evaluation was performed on standard benchmark datasets, which may not fully represent real-world scenarios and challenges.

Future work could focus on exploring techniques to reduce the model's computational demands and improve its efficiency without compromising performance. Additionally, expanding the evaluation to domain-specific datasets and more diverse tasks would provide a broader understanding of the model's capabilities.

In conclusion, the results demonstrate that our large language model outperformed traditional baseline models across multiple benchmark tasks, showcasing its effectiveness in various natural language processing tasks. The statistical analysis confirms the significance and practical significance of the performance improvements. These findings highlight the potential and utility of large language models in advancing natural language understanding and generation tasks.

5. ETHICAL CONSIDERATIONS

5.1 - Data Privacy and Consent:

The benchmark datasets used in this study were collected and publicly released with appropriate consent and permissions from the original data sources. We ensured compliance with data protection regulations and guidelines.

Personally identifiable information (PII) in the datasets was carefully anonymized to protect the privacy of individuals. Any instances of PII were removed or replaced with generic placeholders during the preprocessing phase.

5.2 - Bias and Fairness:

We recognize that biases present in the training data can be inadvertently learned and perpetuated by large language models. To mitigate potential biases, we carefully reviewed the datasets and removed any instances of sensitive or offensive content during the preprocessing stage.

However, it is important to note that biases may still exist in the underlying data sources used to create the benchmark datasets. We encourage future researchers to consider and address potential biases in both data collection and model training processes.

5.3 - Potential for Harmful Content:

Large language models have the potential to generate or amplify harmful and inappropriate content. To mitigate this risk, we implemented content filtering mechanisms during the data preprocessing phase to remove or censor any offensive or harmful content.

We also encourage the responsible use of large language models and urge researchers and practitioners to implement adequate safeguards and moderation mechanisms to prevent the dissemination of harmful or misleading information.

5.4 - User Consent and Control:

As large language models become more integrated into various applications, it is crucial to prioritize user consent and control over the generated outputs. We emphasize the importance of providing users with transparent information about the underlying model and its capabilities, as well as clear options for user consent and opt-out mechanisms.

5.5 - Adherence to Ethical Guidelines:

This study adheres to established ethical guidelines for conducting research in artificial intelligence and natural language processing, including guidelines provided by organizations such as the Partnership on AI and the ACM Code of Ethics.

We also acknowledge the evolving nature of ethical considerations in this field and encourage ongoing discussions and collaborations to address emerging challenges and promote responsible AI practices.

It is important for researchers and practitioners to recognize the ethical implications associated with large language models and to actively work toward mitigating potential risks. By considering and addressing ethical considerations, we can strive for the responsible development and deployment of large language models that benefit society while minimizing harm.

6. CONCLUSION AND FUTURE WORK

6.1 - Conclusion:

In this study, we conducted a performance-based case study of a large language model on various benchmark tasks in natural language processing. Our experimental results demonstrated the superiority of the large language model over traditional baseline models across multiple tasks, including sentiment analysis, named entity recognition, machine translation, and

question-answering. The model's ability to capture contextual information and semantic relationships in text, combined with the fine-tuning process, resulted in significant performance improvements. These findings highlight the potential of large language models in advancing natural language understanding and generation tasks.

6.2 - Contributions:

We provide empirical evidence of the effectiveness and superiority of large language models in various NLP tasks.

Our study contributes to the understanding of the impact of pre-training and fine-tuning on model performance.

The evaluation and comparison with baseline models serve as a benchmark for future research and development in the field of large language models.

6.3 - Future Work:

While this study sheds light on the capabilities and performance of large language models, there are several avenues for future exploration:

Model Optimization: Further research can focus on optimizing large language models to reduce computational requirements and improve efficiency without sacrificing performance. Techniques such as model distillation, quantization, and knowledge distillation can be explored to achieve more lightweight models.

Domain-specific Fine-tuning: Extending the fine-tuning process to domain-specific datasets can help assess the model's effectiveness in specialized domains and tasks. This can involve exploring transfer learning techniques or utilizing task-specific data augmentation strategies.

Bias Mitigation: Addressing biases in large language models is an ongoing challenge. Future work should concentrate on developing techniques to detect and mitigate biases in training data, model behavior, and generated outputs. Collaborations with domain experts and ethicists can provide valuable insights in this area.

User Interface and Control: Enhancing user interfaces and control mechanisms for large language models can empower users to have better visibility, understanding, and control over the model's behavior. This includes providing clear options for user consent, fine-grained customization of model outputs, and transparent explanations of how the model generates responses.

Robustness and Adversarial Attacks: Investigating the robustness of large language models against adversarial attacks and developing defense mechanisms is crucial. Future work can focus on understanding model vulnerabilities, exploring adversarial training methods, and developing techniques to detect and mitigate malicious use cases.

By addressing these areas of future work, we can advance the field of large language models, improve their capabilities, and address potential limitations and ethical concerns.

The ongoing exploration and development of large language models hold tremendous potential for revolutionizing natural language processing applications and enabling more sophisticated and context-aware language understanding and generation systems.

REFERENCES

[1] [Devlin et al., 2019] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171-4186).

[2] [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[3] [Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL: https://openai.com/blog/language-unsupervised/.

[4] [Brown et al., 2020a] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[5] [Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

[6] [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.

N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[7] [Brown et al., 2020b] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[8] [Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems (pp. 5754-5764).

[9] [Lewis et al., 2020] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871-7880).

[10] [Liu et al., 2019] Liu, Y., Khandelwal, U., Levy, O., Lewis, M., & Zettlemoyer, L. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[11] [Lan et al., 2020] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[12] [Wang et al., 2019] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2019 International Conference on Learning Representations (ICLR).

[13] [Lample & Conneau, 2019] Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

[14] [Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[15] [Lewis et al., 2020] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871-7880).

[16] [Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[17] [Lan et al., 2020] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[18] [Sun et al., 2020] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, Y., Wu, H., ... & Zhou, M. (2020). ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2004.09977.

[19] [Beltagy et al., 2020] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[20] [Liu et al., 2019] Liu, Y., Khandelwal, U., Levy, O., Lewis, M., & Zettlemoyer, L. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[21] [Zhao et al., 2023] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . .Wen, J. (2023). A Survey of Large Language Models. *ArXiv*. /abs/2303.18223.

[22] [Vaswani et al., 2017] (updated version 2023) Attention Is All You Need http://arxiv.org/abs/1706.03762.